JURNAL MEDIA APLIKOM
ISSN : 2086 - 972X
Vol. 2, No. 2, Mei 2012

BATAK KARO ALPHABET PATTERN RECOGNITION

Oskar Ika Adi Nugroho
Information System Departement
STIKOM Yos Sudarso Purwokerto

Abstract

Karo Batak script is typical and ancestral heritage of Batak tribe to preserve its existence. Currently Batak Karo script already has a standard of Unicode. Therefore, the authors develop models and handwriting recognition software Batak Karo script in real time. Real time or online is the way the introduction of OCR (Optical CharacterRecognize). Batak script used is Batak Karo alphabet has 20 types of characters. Model and software that is built using feature extraction zoning and consider the long strokes. The algorithm used is the Kohonen algorithm neuralnetwork. Kohonen neural network, including unsupervised learning (learning uncontrolled) by studying the distribution of the set of patterns without any class information.

Keywords: *Batak Karo Alphabet, Pattern Recognition,Kohonen Network Model, Unsupervised learning, OCR—Online*

1. Introduction

The recognition of characters and handwriting has been a popular research area for many years because of its potentials and various applications, such as postal automation, bank cheque processing, documents analysis, reading postal codes and reading different forms, etc. Several scientific researchers have been carried out for handwritten recognition of English language, Chinese/Japanese/Hindi languages, and Arabic/ Farsi language .[7]
This paper will discuss how pattern recognition Batak karo characters into Latin characters using artificial neural network model with Kohonen networks. The number of Batak karo characters selected are 21 characters.
The paper is organized as follows: Section 2 presents an overview of Karo Batak language. Section 3 display system architecture of the proposed method and discussion of each, ways of processing and features. Section 4, experimental results and section 5, conclusion.

2. The Batak Karo Language

Batak is a collective term used to identify a number of ethnic groups predominantly found in North Sumatra, Indonesia. The term is used to include the Toba, Karo, Pakpak, Simalungun, Angkola and Mandailing, each of which are distinct but related groups with distinct, albeit related, languages and customs (*adat*). Occasionally it is also used to include the Alas people of Central/Southern Aceh, but usually only as relates to language groups. **Batak Karo**, referred to in Indonesia simply as **Bahasa Karo** (Karo language), is an Austronesian language that is spoken by the Karo people of Indonesia. It is used by around 600,000 people in North Sumatra. It was historically written using the Batak alphabet which is descended from the Brahmi script of ancient India by way of the Pallava and Old Kawi scripts.

Batak script originally only be understood and understood by only a very limited circle of experts mejik (magic) and treatment (datu or teacher). So pustaha generally written by the datu. Group of religious leaders (parbaringin and Parmalim) also understand it but just write things religious ordinances, because they did not and does not interfere in the affairs datu mejik. Pustaha it mostly contains about kedukunan, drugs, and astrologer. So in pustaha not found on genealogy (tarombo), literature, poem, poetry (turi-turian, umpasa) which is derived from generation to generation orally. but nowadays only a tiny number of Karo can write or Understand the script, and instead the Latin alphabet is used. [2],[3]

Kohonen or Self Organizing Map (SOM) was first introduced by Teuvo Kohonen in the early 1980's. Kohonen unsupervised learning method to learn to do including the set of distribution patterns without any class information.[1],[4],[5]

Code	Batak Karo Characters	Correspondents in Latin Characters
1BC0	ᯀ	HA
1BC2	ᯂ	KA
1BC6	ᯆ	BA
1BC7	ᯇ	PA
1BC9	ᯉ	NA
1BCB	ᯋ	WA
1BCE	ᯎ	GA
1BD0	ᯐ	JA
1BD1	ᯑ	DA
1BD2	ᯒ	RA
1BD4	ᯔ	MA
1BD7	ᯗ	TA
1BD8	ᯘ	SA
1BDB	ᯛ	YA
1BDD	ᯝ	NGA
1BDE	ᯞ	LA
1BE1	ᯡ	NDA
1BC5	ᯅ	MBA

1BE4		I
1BE5		U

Table 1. Batak Karo Characters with their Correspondents in Latin Characters.

3. The Method and System Architecture

3.1. *OCR (Optical Character Recognition)*

Optical Character Recognition (OCR) is a system that can recognize good writing it on paper document (type, scan) as well as handwriting. To recognize the writing, OCR have two ways, namely off-line and on-line. Off-line handwriting recognition is a way to input an image's scan results. While on-line handwriting recognition is a way to recognize handwriting input's direct or strokes in the form of graffiti writing that was written in real time on writing digitalmedia. In general, stages of handwriting recognition OCR on-line system can be seen in the figure below:

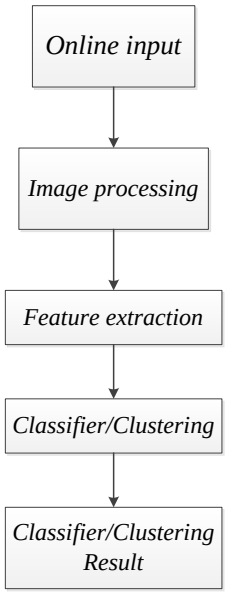


Figure 3.1: Stages of handwriting recognition

3.2. Image Processing

Image processing has several types to process the image, but in this study are used in image processing such as segmentation and scaling.

3.2.1. Segmentation

Aiming to download segments or capture the image in order to do the cutting process images [11]. Segmentation process is done by searching the border of the image that want to

capture, ranging from the upper limit, lower limit, limit the right and the left boundary. Segmentation process begins by scanning the image from left to right and from top to bottom. Scan process is to determine the initial presence of a black pixel to the next meeting on the last black pixel in the image. After the scan is complete, then do the boxing or restriction area an image that has a black pixel.

3.2.2. Scalling

Scaling is the process to change the size of an image [11]. Scaling is the process of normalizing the size so obtained is always the same size although the size of the text or images are not the same (large or small). In this study the results of scaling the image size is set to the dimensions 5x7.

.

3.3.Feature Extraction

Feature extraction aims to obtain the characteristics of a character which distinguishes it from other characters, called the feature [12]. In this research, feature extraction used is zoning. Zoning is one of the Feature Extraction of statistical type feature.Metode zoning is to divide the characters into N x M region. From each region, features extracted to form a feature vector. The goal is to obtain zoning local characteristics in addition to the global characteristics. To obtain the characteristics of a character, is also used to calculate the length of stroke. The results of calculating the length of stroke is a decimal number that is converted to binary (0 and 1) Here is an example of the results of feature extraction (zoning and compute the length of stroke :

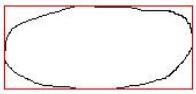
Stroke	Feature Extraction	Result
		11111100011000110001100011000111111
		11111100010000101110110100000001110

Table 2. Example of the results of feature extraction

3.4. Kohonen or Self Organizing Map (SOM)

Kohonen or Self Organizing Map (SOM) was first introduced by Teuvo Kohonen in the early 1980's. In Kohonen's algorithm a two-dimensional array of output nodes is used to form feature maps. Every input is connected to every output node through a variable connection weight, and these output nodes are extensively interconnected with many local connections, as shown in Fig. 3.1. These weights (from input node *i* to output node *j*) will be organized such that topologically close nodes are sensitive to inputs that are physically similar. Output nodes will thus be ordered in a natural way. The low-level organization in this feature map is

generally predetermined, while some of the organization at higher levels is created during learning by algorithms that promote self-organization.

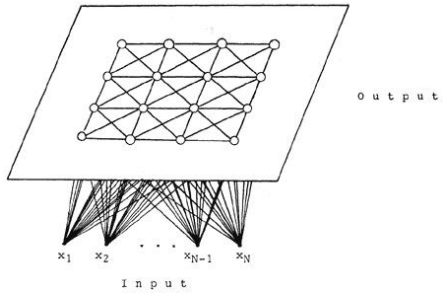


Figure 3.2: Kohonen's self-organizing feature maps.

In brief, the self-organizing feature maps can be generated by the following steps:

1. To start with, set these weights from inputs (say, N in number) to outputs (m in number) as small random values.
2. Present inputs to the system to perform the clustering.
3. Compute the distances dj between the inputs and each output node j by using the following expression:

$$d_j = \sum_{i=1}^{N} \left[x_i(t) - W_{ij}(t) \right]^2$$

where $x_i(t)$, $i = 1, 2, \ldots, N$, is the input to node i at time t, and $W_{ij}(t)$ is the weight from input node i to output node j at time t.

4. Use this distance measure to find the cluster, that is, to select the output node with minimum d_j.
5. Update these weights through the following iteration:

$$W_{ij}(t+1) = W_{ij}(t) + \eta(t)\left[x_i(t) - W_{ij}(t) \right]$$
$$i = 1, 2, \ldots, N$$

where η (t) is a gain term and decreases with time. It ranges from 0 to 1.

From the procedure described above, it can be seen that this process of feature map forming nets is similar to the K-means clustering algorithm. No information concerning the correct class is needed during adaptive training.

4. Experimental Result

Tests carried out by using the device software that has been created with the following scenario:

1. Tests to determine the accuracy of handwriting recognition Karo Batak script by hand using Kohonen neural network algorithm.

2. Tests to determine the accuracy of handwriting recognitionKaro Batak script by hand using Kohonen neural network algorithm with noise.

References

[1] Teuvo Kohonen, "Self-Organizing Maps", 3rd edition,,Springer Series in Information Sciences 30, Springer-Verlag, Berlin Heildelberg New York, 2001.

[2] Uli Kozok, Warisan Leluhur-Sastra Lama dan Aksara Batak, KPG, Jakarta, 1999.

[3] Uli Kozok, The Seal of the last Singamangaraja, Indonesia and the Malay World, Vol 28 No 82, Carfax Publishing, 2000.

[4] Heaton, Jeff, Introduction to Neural Networks with Java, Second Edition, Heaton Research, Inc, Chesterfield, 2008.

[5] Heaton, Jeff, Introduction to Neural Networks for C#, Second Edition, Heaton Research, Inc Chesterfield,, 2008.

[6] Mubarok, Lala Septem Riza MT., Dr. Wawan Setiawan M.Kom., "*Pengenalan Tulisan Aksara Sunda Menggunakan Kohonen Neural Network"*, Universitas Pendidikan Indonesia, Bandung, 2011.

[7] Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass, *"Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character"*, International Journal of Advanced Science and Technology Vol. 33, SERSC ,August 2011.

[8] Bow, Sing-Tze., Pattern recognition and image preprocessing, MARCEL DEKKER, INC., New York, 1992.

[9] Mohamed Cheriet ... [et al.]., Character recognition systems : a guide for students and practioners, John Wiley & Sons, Inc., New York, 2007.

[10] Shunji Mori, Hirobumi Nishida, Hiromitsu Yamada., Optical character recognition, John Wiley & Sons. Inc., New York, 1999.

[11] Asworo. Perbandingan antara metode kohonen Neurall network dan learning vector Quantization pada sistem pengenalan Tulisan tangan secara real time, Institut Teknologi Sepuluh November. Surabaya, 2010.

[12] Kusumadewi, Sri.. Pengenalan Artificial Neural Network, Handout kuliah, Institut Teknologi Bandung, Bandung:, 2009.